# 电子 AI+系列专题报告

## 边缘 AI：大语言模型的终端部署，推动新一轮终端需求

超配

## 核心观点

**大模型参数量级飞涨，相应训练集需同比提升。**李开复定义 AI 2.0 时代的特征是通过海量数据，无需标注自监督学习，训练一个基础大模型，并在各领域将其专业化。据相关论文，当模型的参数量大于某阈值，会展现出类似推理、无监督学习等未曾出现的能力，这种现象被称为“涌现”，因此目前大语言模型参数均在十亿量级以上。同时，Deepmind 研究表明，模型参数的上涨需要配合等比例上升的优质数据集来达到最佳训练效果。因此，大模型参数在十亿级以上发展并受限于优质数据集的增速是 AI 发展的必然趋势。

**大模型增长挑战芯片算力和内存，无法实现完整端侧部署。**大模型训练和推理的三大瓶颈是算力、显存和通信，根据我们的测算，算力方面 GPT-3 训练所需算力为 121528 TFLOPS，若 30 天内完成，需要 1558 颗 A100。内存角度，GPT-3 训练至少需要 3.2T 内存，至少 44 张 A100，推理任务则主要受显存限制，需要 4 至 8 张 A100，因此完整的模型无法在终端上离线运行。

**优化后大模型可在旗舰机型芯片上运行，AI 落地有望推动新一轮换机潮。**AI 部署本地化具有必要性，优势包括更低的延迟、更小的带宽、提高数据安全、保护数据隐私、高可靠性等。完整的大模型仅参数权重就占满一张 80G 的 GPU，但是通过量化、知识蒸馏、剪枝等优化，大模型可以在手机本地实现推理。高通团队使用骁龙 8 Gen2 部署 Stable Diffusion，实现本地运营 15 秒出图，证明了大模型本地化运行的可能，也体现出目前手机芯片的局限性。根据 IDC 数据，1Q23 全球手机销量中主处理器频率超过 2.8GHz 的占比 36%，价格在 1000 美金以上的占比 13%，即旗舰机型占比较低，随着 AI 大模型在边缘端落地，有望推动新一轮换机潮。

**以大语言模型为核心，以语言为接口，控制多 AI 模型系统，构建“贾维斯”式智能管家。**我们认为大语言模型不仅可以实现对话、创意，未来也有望作为众多复杂 AI 模型的控制中心，同时也是接受用户指令的交互窗口，实现《钢铁侠》电影中“贾维斯”式综合智能管家。23 年 5 月，Google 推出 PaLM 2 轻量版 Gecko，其可在最新的旗舰机型上离线运行。同月，OpenAI 首次推出 ChatGPT 移动端应用，各家大厂正式进入 AI 模型移动端创新、竞争时期。智能音箱、全屋智能中控屏、手机、MR 等均有望成为这一时代的交互入口。

**产业链相关公司：**半导体：晶晨股份、瑞芯微、全志科技、北京君正、兆易创新；消费电子：传音控股、歌尔股份、福立旺、闻泰科技、创维数字。

**风险提示：**AI 技术发展不及预期；边缘端芯片发展不及预期。

## 重点公司盈利预测及投资评级

| 公司代码 | 公司名称 | 投资评级 | 昨收盘（元） | 总市值（亿元） | EPS 2023E | EPS 2024E | PE 2023E | PE 2024E |
|---|---|---|---|---|---|---|---|---|
| 688099.SH | 晶晨股份 | 买入 | 86.36 | 358.62 | 1.77 | 2.23 | 48.79 | 38.66 |
| 300223.SZ | 北京君正 | 买入 | 93.77 | 451.57 | 1.64 | 1.87 | 57.22 | 50.23 |
| 688036.SH | 传音控股 | 买入 | 127.95 | 1,028.65 | 3.09 | 4.61 | 41.41 | 27.73 |
| 002241.SZ | 歌尔股份 | 买入 | 18.29 | 625.59 | 0.52 | 0.76 | 35.17 | 24.22 |
| 688678.SH | 福立旺 | 买入 | 18.70 | 32.42 | 0.94 | 1.35 | 19.89 | 13.89 |
| 600745.SH | 闻泰科技 | 买入 | 50.08 | 622.40 | 0.94 | 1.35 | 53.28 | 37.20 |
| 000810.SZ | 创维数字 | 买入 | 15.74 | 181.04 | 0.94 | 1.35 | 16.74 | 11.69 |

资料来源：Wind、国信证券经济研究所预测

行业研究·行业专题

电子

超配·维持评级

**证券分析师：胡剑**
021-60893306
hujian1@guosen.com.cn
S0980521080001

**证券分析师：胡慧**
021-60871321
huhui2@guosen.com.cn
S0980521080002

**证券分析师：周靖翔**
021-60375402
zhoujingxiang@guosen.com.cn
S0980522100001

**证券分析师：李梓澎**
0755-81981181
lizipeng@guosen.com.cn
S0980522090001

**证券分析师：叶子**
0755-81982153
yezi3@guosen.com.cn
S0980522100003

**联系人：詹浏洋**
010-88005307
zhanliuyang@guosen.com.cn

**联系人：李书颖**
0755-81982362
lishuying@guosen.com.cn

市场走势


资料来源：Wind、国信证券经济研究所整理

相关研究报告

《电子行业周报-AI+开启半导体新周期》——2023-05-29
《复盘英伟达的 AI 发展之路》——2023-05-29
《电子行业周报-半导体周期拐点临近，国产化进程提速》——2023-05-24
《电子行业周报-景气拐点将至，以时间换空间》——2023-05-15
《电子行业周报-在行业周期筑底阶段无需过度悲观》——2023-05-08

# 内容目录

# 图表目录

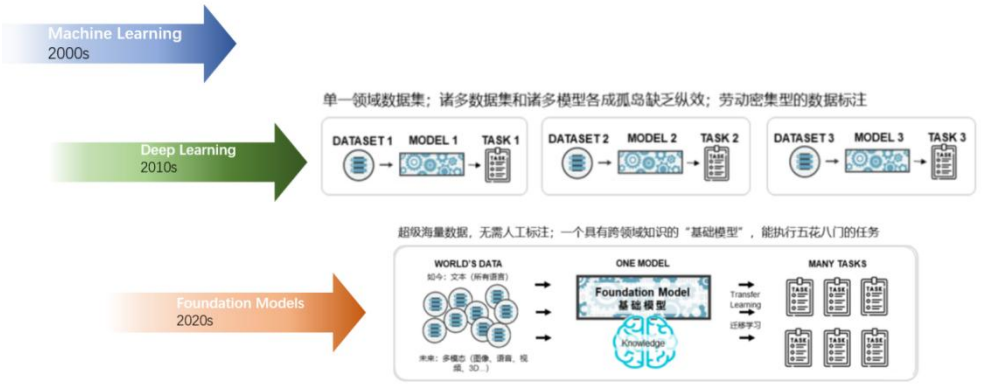
# 百亿参数大模型具备涌现能力，训练数据需等比例提升

## 大模型的参数下限：AI2.0 时代，基础大模型参数指数级增长

**李开复提出本次由 GPT-4、ChatGPT 引发的 AI 新机遇与之前有所不同，属于 AI 2.0 时代。**AI 1.0 时代具体指的是以 CNN（卷积神经网络）为核心，机器视觉和自然语言处理快速发展的时期，暴涨的数据量伴随搜集、清洗、标注整个过程的成本增加，且单一领域的数据集和模型形成孤岛，每个领域和应用的优化都是割裂的，难以形成“通用”。

**AI 2.0 时代的特征是通过海量数据，无需标注自监督学习，训练一个基础大模型，并在各个应用领域将其专业化。**具体来说有三个特点：1）对于拥有的超级海量的数据，无需进行人工标注，即进行自监督学习；2）基础模型规模非常大，参数规模从十亿到千亿级别；3）训练出的基础模型具有跨领域知识，而后通过微调用降低成本的方法来训练，以适应不同领域的任务。AI 2.0 的巨大跃迁之处在于，它克服了前者单领域、多模型的限制。

图1：AI2.0 时代的特征是通过超级海量数据无需标注训练一个大模型

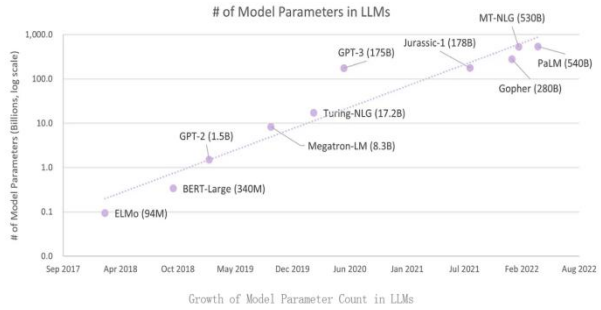



资料来源：创新工场，国信证券经济研究所整理

以大语言模型（Large Language Model，LLM）为例，语言模型已经存在了几十年，从最基本的 N-gram 模型（语言由简单的向量表示），到更复杂的 RNN 模型、LSTM 神经网络，再到 2017 年 Google Brain 提出 Transformer。Transformer 不再基于对每个单词的单独理解进行处理，而是将句子和段落作为一个整体进行处理，使 LLM 能够从自然语言中深入理解人类的意图，并让一系列应用成为可能：从描述中生成艺术创作、将大量非结构化数据提炼成简洁的摘要、更准确的翻译、回答复杂的查询等。

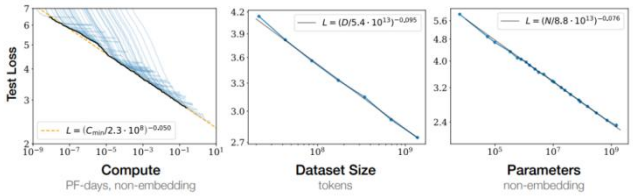
**以模型中的参数数量衡量，大型语言模型的参数在过去五年中以指数级增长。**模型的性能非常依赖于模型的规模，具体包括：参数数量、数据集大小和计算量，模型的效果会随着三者的指数增加而线性提高，这种现象被称为 Scaling Law（缩放能力）。

图2：过去五年 LLM 模型参数快速增长




资料来源：Sunyan’s Substack，国信证券经济研究所整理

图3：参数量的指数提升线性提高模型性能

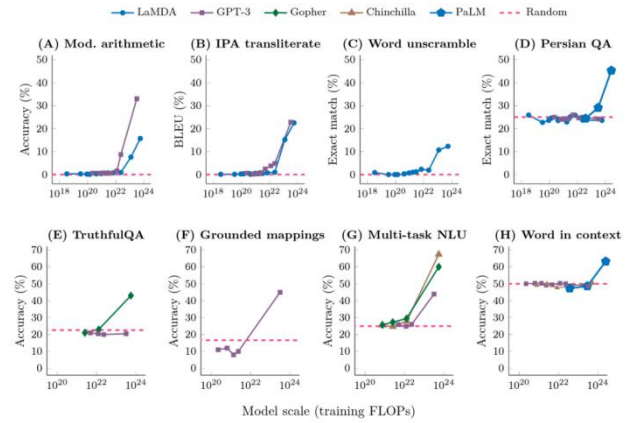



资料来源：《Scaling laws for neural language models》，国信证券经济研究所整理

**当模型的参数量大于一定程度的时候，模型能力会突然提升，并拥有一些未曾出现的能力，如推理能力、无标注学习能力等，这种现象被称为涌现能力。**在 Jason Wei 的论文中，具体定义为“在小模型中没有表现出来，但是在大模型中变现出来的能力”。“涌现能力”只是对一种现象的描述，而并非模型的某种真正的性质，出现涌现能力的原因也尚待探索。

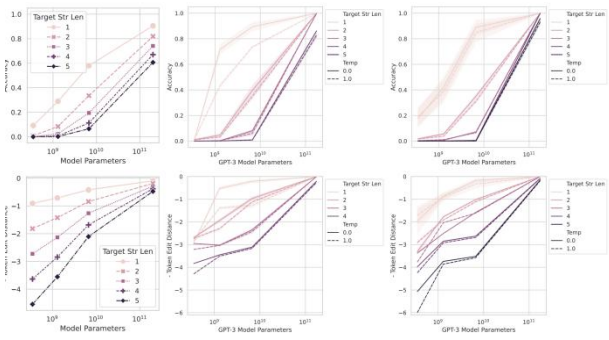
其中一种观点认为，大模型的涌现能力来自于其不连续的评价指标。如果换成更为平滑的指标，相对较小的模型的效果也并非停滞不前，规模在阈值以下的模型，随着规模的提高，生成的内容也在逐渐靠近正确答案。斯坦福的研究人员将 NLP 中不连续的非线性评价指标转为连续的线性评价指标，结果模型的性能变得更加平滑、可预测。**具体来看，$10^9$以上模型能力提升加速，因此目前来看 $10^9$几乎是大语言模型参数量的下限。**

图4：当模型的参数量大于一定程度时模型效果会突然提升




资料来源：《Language models are few-shot learners》，国信证券经济研究所整理

图5：小模型的性能也随着规模扩大而逐步提高




资料来源：《Are Emergent Abilities of Large Language Models a Mirage?》，国信证券经济研究所整理

## 大模型的参数上限：参数的增加需要同等量级的训练集增加

**参数数量的增速应与训练 token 数量的增长速度大致相同，从而让模型损失(L)实现最小化，性能最大化。**Deepmind 在《Traning Compute-Optimal Large Language Models》中，通过在 5 到 5000 亿个 token 上训练 400 多个语言模型，参数个数范

围从 7000 万到 160 亿，发现模型大小和训练集数量应该相等地缩放，从而达到最佳效果。目前看来，单一语言模态的大模型，100B 量级的参数足以满足大多数知识检索和浅层推理的需求，但充分释放这些参数的全部潜力需要 1000B 量级的训练 token。

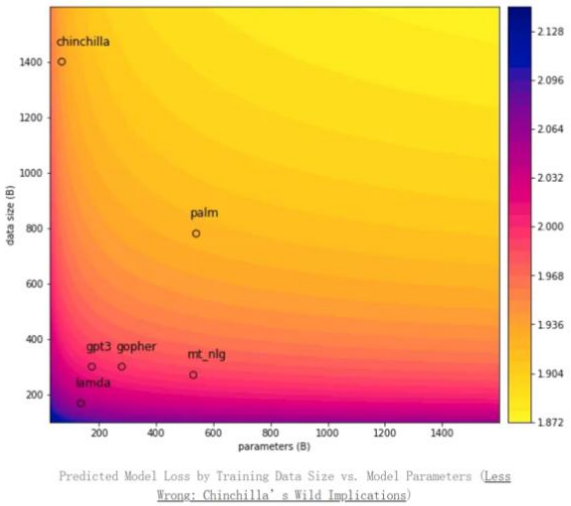
作为验证，通过训练一个预测的计算最优模型 Chinchilla 来检验这个假设，该模型使用与 Gopher 使用相同的 FLOTs，但具有 70B 个参数和 4 倍多的数据，最终在大量下游评估任务中，Chinchilla 表现显着优于 Gopher，且其缩小的模型尺寸大大降低了推理成本，并极大地促进了下游在较小硬件上的使用。

图6：2022 年最大的五个 transfomer 模型条件

| Model | Size (# Parameters) | Training Tokens |
|---|---|---|
| LaMDA (Thoppilan et al., 2022) | 137 Billion | 168 Billion |
| GPT-3 (Brown et al., 2020) | 175 Billion | 300 Billion |
| Jurassic (Lieber et al., 2021) | 178 Billion | 300 Billion |
| *Gopher* (Rae et al., 2021) | 280 Billion | 300 Billion |
| MT-NLG 530B (Smith et al., 2022) | 530 Billion | 270 Billion |
| *Chinchilla* | 70 Billion | 1.4 Trillion |

资料来源：DeepMind，国信证券经济研究所整理

图7：各模型位于 LM 损失等高线图上的位置

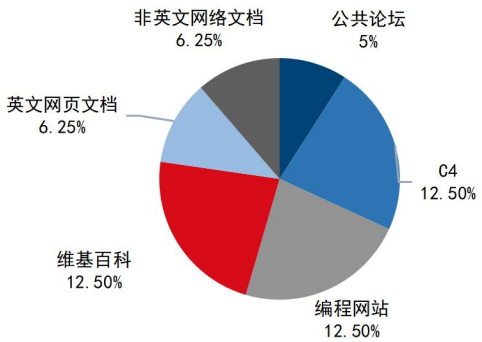



Predicted Model Loss by Training Data Size vs. Model Parameters (Less Wrong: Chinchilla's Wild Implications)

资料来源：Sunyan's Substack，国信证券经济研究所整理

**因此，优质大模型的训练，高质量的大数据集是必要条件。**目前主要的数据获取渠道是公开的论坛，例如谷歌的 LaMDA 模型，在论文中表示其预训练数据 50%对话数据来自公共论坛；12.5% C4 数据；12.5%的代码文档来自与编程相关的网站；12.5%维基百科；6.25%英文网页文档；6.25%的非英语网络文档，数据集中的单词总数为 1.56T，而 OpenAI 使用了 45T 数据。未来如何获得高质量的训练集始终是各家大厂的首要竞争领域。

图8：LaMDA 模型训练数据来源




资料来源：谷歌，国信证券经济研究所整理

表1：GPT 参数和训练集规模快速增长

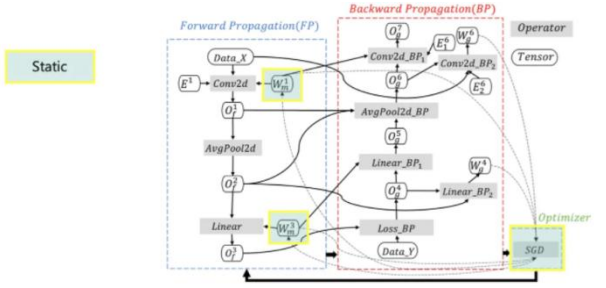
| 模型 | 发布时间 | 参数量 | 数据规模 | Tokens |
|---|---|---|---|---|
| **GPT** | 2017.6 | 1.17 亿 | 5GB | 1.17 亿 |
| **GPT2** | 2019.2 | 小：1.24 亿<br>中：3.55 亿<br>大：7.74 亿<br>超大：15 亿 | 40GB | 15 亿 |
| **T5** | 2019 | 小：0.6 亿<br>基础：2.2 亿<br>大：7.7 亿<br>TB-3B：30 亿<br>T5-11B：110 亿 | 50G | 340 亿 |
| **GPT3** | 2020.6 | 1750 亿 | 45TB | 1750 亿 |
| **ChatGPT** | 2020.6 | 1750 亿 | >45TB | 7740 亿 |

资料来源：ChatGPT，Google，国信证券经济研究所整理

## 大模型训练对硬件的挑战：算力、内存和通信

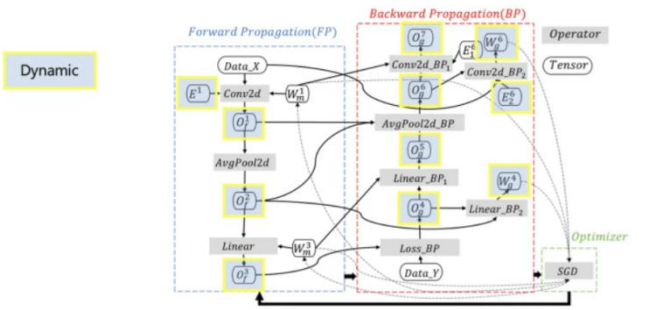
**内存方面，大模型训练的内存可以大致理解为参数、优化器状态、激活、梯度四部分的和。**它们大致分为两类：静态内存和动态内存。参数、优化器状态较为固定，属于静态内存，激活和梯度等中间变量属于动态内存，是最主要的内存占用原因，动态内存通常是静态内存的数倍。

图9：静态内存




资料来源：知乎，国信证券经济研究所整理

图10：动态内存




资料来源：知乎，国信证券经济研究所整理

**我们可以粗略的计算训练 1750 亿参数的 GPT3 所需内存，大约需要 3.2TB 以上。**静态内存方面，大多数 Transformer 都是以混合精度训练的，如 FP16+FP32，以减少训练模型内存，则一个参数占 2 个字节，参数和优化器状态合计占用内存 1635G。而动态内存，根据不同的批量大小、并行技术等结果相差较大，通常是静态内存的数倍。更简洁的估算方法，可以假设典型的 LLM 训练中，优化器状态、梯度和参数所需的内存为 20N 字节，其中 N 是模型参数数量，则 1750 亿参数的 GPT3 大概需要 3.2TB 内存。

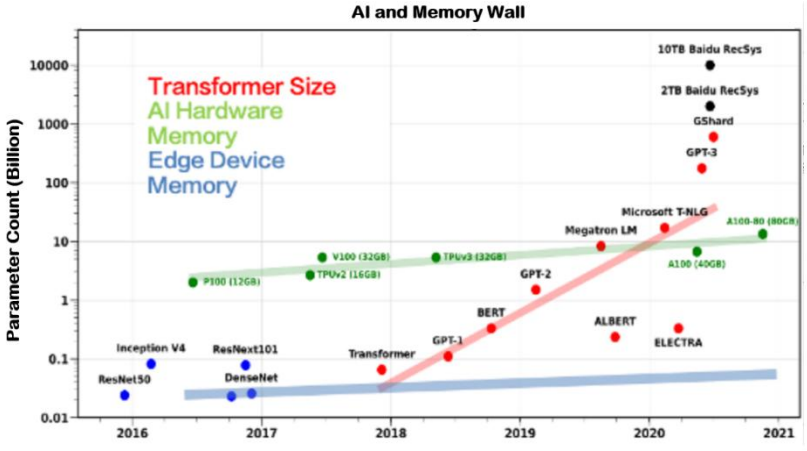
推理所需内存则较小，假设以 FP16 存储，175B 参数的 GPT3 推理大约需要内存 327G，则对应 4 张 80G A100，如果以 FP32 运算，则需要 10 张。

表2：大语言模型的计算

| | 公式 | 注释 |
|---|---|---|
| **模型参数** | • 混合精度（fp16/bf16 和 fp32），$\text{memory}_{\text{model}} = (2 \text{ bytes/param}) \cdot (\text{No. params})$<br>• 对于vanilla AdamW，$\text{memory}_{\text{optimizer}} = (12 \text{ bytes/param}) \cdot (\text{No. params})$<br>▪ fp32参数副本：4字节/参数<br>▪ 动量：4字节/参数<br>▪ 方差：4字节/参数 | |
| **优化器内存** | • 对于像bitsandbytes这样的8位优化器，$\text{memory}_{\text{optimizer}} = (6 \text{ bytes/param}) \cdot (\text{No. params})$<br>▪ fp32参数副本：4字节/参数<br>▪ 动量：1字节/参数<br>▪ 方差：1字节/参数<br>• 对于具有动量的类SGD优化器，$\text{memory}_{\text{optimizer}} = (8 \text{ bytes/param}) \cdot (\text{No. params})$<br>▪ fp32参数副本：4字节/参数<br>▪ 动量：4字节/参数 | |
| **梯度内存** | • fp32，$\text{memory}_{\text{gradients}} = (4 \text{ bytes/param}) \cdot (\text{No. params})$<br>• fp16，$\text{memory}_{\text{gradients}} = (2 \text{ bytes/param}) \cdot (\text{No. params})$ | |
| **激活重计算** | $\text{memory}_{\text{activations}}^{\text{No Recomputation}} = sbhL(10 + \frac{24}{t} + 5\frac{a \cdot s}{h \cdot t})$ bytes<br>$\text{memory}_{\text{activations}}^{\text{Selective Recomputation}} = sbhL(10 + \frac{24}{t})$ bytes<br>$\text{memory}_{\text{activations}}^{\text{Full Recomputation}} = 2 \cdot sbhL$ bytes | $a$ number of attention heads; $b$ microbatch size; $h$ hidden dimension size; $L$ number of transformer layers; $p$ pipeline parallel size; $s$ sequence length; $t$ tensor parallel size; $v$ vocabulary size |
| **模型训练内存需求** | $\text{Total Memory}_{\text{Training}} = \text{Model Memory} + \text{Optimiser Memory} + \text{Activation Memory} + \text{Gradient Memory}$ | |
| **模型推理内存需求** | $\text{Total Memory}_{\text{Inference}} \approx (1.2) \times \text{Model Memory}$ | |

资料来源：Eleutherai，国信证券经济研究所整理

图11：模型大小与设备内存的增长示意图




资料来源：NVIDIA，国信证券经济研究所整理

**算力方面，**根据 OpenAI 在 2020 年发表的论文，训练阶段算力需求是模型参数数量与训练数据集规模乘积的 6 倍：训练阶段算力需求=6×模型参数数量×训练集规模；推理阶段算力需求是模型参数数量与训练数据集规模乘积的 2 倍：推理阶段算力需求=2×模型参数数量×训练及规模。

**训练阶段：**考虑采用精度为 32 位的单精度浮点数数据进行训练和推理。以 A100 PCIe 芯片为例（H100 PCIe 芯片同理），根据前述公式，GPT-3 训练所需运算次数为：样本

token 数 3000 亿个*6*参数量 1750 亿个=315*10^21FLOPs；考虑训练时间要求在 30 天完成（训练时间为 2592000 秒），则对应 GPT-3 训练所需算力为 121528TFLOPS；结合 A100 有效算力 78TFLOPS，得到所需 GPU 数量为 1558 个，对应 AI 服务器为 195 台。

**推理阶段：**按谷歌每日搜索量 35 亿次进行估计，假设每次访问提问 4 次，每次提问+回答需处理字数 425 字，平均每个字转换为 token 比例为 4/3，则每日 GPT-3 需推理 token 数为 79330 亿个，则推理所需运算次数为 4760*10^21FLOPs；考虑推理时间以每日为单位（推理时间为 86400 秒），则对应 GPT-3 推理所需算力为 55*10^6TFLOPS；结合 A100 有效算力 78TFLOPS，得到所需 GPU 数量为 706315 个，对应 AI 服务器为 8.8 万台。

图12：算力计算公式


资料来源：NVIDIA，国信证券经济研究所整理

图13：近年推出的大预言模型有效算力比率

| 模型名称 | 推出时间 | 使用硬件 | 有效算力比率 |
|---|---|---|---|
| GPT-3 | 2020 年 5 月 | 英伟达 V100 | 21.3% |
| MT-NLG | 2021 年 10 月 | 英伟达 A100 | 30.2% |
| PaLM | 2022 年 4 月 | 谷歌 TPU | 46.2% |

资料来源：NVIDIA，国信证券经济研究所整理

表3：大预言模型算力测算

| | | A100 PCIe | H100 PCIe |
|---|---|---|---|
| 训练阶段算力需求测算 | Tensor Float 32(TF32) | 156TFLOPS | 756TFLOPS |
| | 有效算力 | 78TFLOPS | 378TFLOPS |
| | GPT-3 训练所需运算次数 | 315*10^21FLOPs | 315*10^21FLOPs |
| | GPT-3 训练所需算力 | 121528TFLOPS | 121528TFLOPS |
| | 所需 GPU 数量 | 1558 | 322 |
| | GPU 单价 | 1.5 万美元 | 3.65 万美元 |
| | 对应 GPU 价值 | 2337 万美元 | 1175.3 万美元 |
| 推理阶段算力需求测算 | Tensor Float 32(TF32) | 156TFLOPS | 756TFLOPS |
| | 有效算力 | 78TFLOPS | 378TFLOPS |
| | GPT-3 推理所需运算次数 | 4760*10^21FLOPs | 4760*10^21FLOPs |
| | GPT-3 推理所需算力 | 55*10^6TFLOPS | 55*10^6TFLOPS |
| | 所需 GPU 数量 | 706315 | 145748 |
| | GPU 单价 | 1.5 万美元 | 3.65 万美元 |
| | 对应 GPU 价值 | 105.95 亿美元 | 53.2 亿美元 |

资料来源：NVIDIA，国信证券经济研究所整理

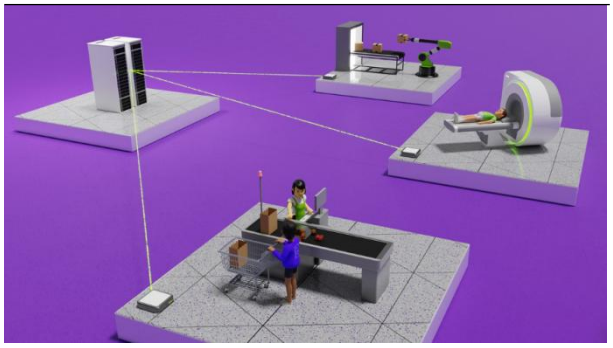
**因此，训练大模型必然需要采用分布式方案。**不仅要满足算力的需求，还要解决上千块 GPU 的分布式训练问题，需要考虑到上百台服务器之间的通信、拓扑、模型并行、流水并行等，这也是复现 GPT-3 的核心难点，模型发布后一年也只有 NVIDIA、微软等大厂成功复现，目前开源的 GPT 模型库就主要是 NVIDIA 的 Megatron-LM 和微软的 DeepSpeed。

# 终端部署具有必要性，轻量化技术优化模型

## 超低时延的智慧场景，终端部署具有必要性

云计算和边缘计算的主要区别在于处理所在的位置。边缘计算，处理发生在网络边缘，更靠近数据源，而云计算，处理发生在数据中心。**边缘计算是指在尽可能靠近数据源或终端的地方捕获和处理数据。**通过在数据源的物理位置附近放置服务器或其他硬件来处理数据，在本地完成处理而不是在云端或集中式数据中心，它能最大限度地减少延迟和数据传输成本，允许实时反馈和决策。

图14：边缘计算的应用场景



资料来源：NVIDIA，国信证券经济研究所整理

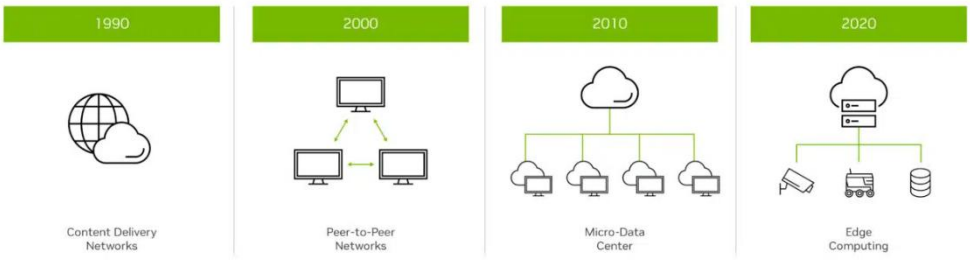
图15：云计算与边缘计算的区别

| Cloud Computing | Edge Computing |
|---|---|
| Non-time-sensitive data processing | Real-time data processing |
| Reliable internet connection | Remote locations with limited or no internet connectivity |
| Dynamic workloads | Large datasets that are too costly to send to the cloud |
| Data in cloud storage | Highly sensitive data and strict data laws |

资料来源：NVIDIA，国信证券经济研究所整理

边缘计算的历史可以追溯到上世纪 90 年代，当时内容分发网络（CDN）充当分布式数据中心。但 CDN 仅限于缓存图像和视频，而不是海量数据工作负载；2000 年左右，智能设备的爆炸式增长给现有 IT 基础设施带来了压力，诸如点对点(P2P)网络的发明减轻了这种压力，在这种网络中，计算机无需通过单独的集中式服务器计算机即可连接并共享资源；10 年代，大公司开始通过公共云向终端用户出租计算和数据存储资源；2020 年后，边缘计算融合了 CDN 的低延迟能力、P2P 网络去中心化平台以及云的可扩展性和弹性，共同构建了一个更高效、更有弹性和更可靠的计算框架。

图16：云计算与边缘计算




资料来源：NVIDIA，国信证券经济研究所整理

目前，越来越多的场景将计算基础设施更靠近传入数据源，让 AI 模型在云端训练，并部署在终端设备上。例如计算机视觉等高度数据密集型、低时延要求类的任务，将 AI 模型部署在终端的优势包括：

**1）更低的延迟：**因为传感器和物联网设备产生的数据不再需要发送到集中式云进

行处理，可以实现更快的响应，获得结果的时间可能从几秒减少到几分之一秒。

**2）减少带宽：**当数据发送到云端时，它通过广域网传输，需要满足全球覆盖和高带宽，成本较高。而边缘计算可以利用局域网处理数据，从而以更低的成本获得更高的带宽。

**3）数据安全：**边缘计算允许组织将所有数据和计算保存在合适的位置，关键数据不需要跨系统传输，减少遭受网络安全攻击的风险。

**4）保护用户隐私：**人工智能可以分析现实世界的信息，而无需将其暴露给人类，大大增加了任何需要分析外貌、声音、医学图像或任何其他个人信息的隐私安全。即使部分数据是出于培训目的而上传，也可以将其匿名化以保护用户身份。

**5）高可靠性：**去中心化和离线功能使边缘 AI 更加稳定，不受网络访问限制，这是关键任务系统稳定运行的必要条件。

当边缘 AI 应用程序遇到它无法准确处理的数据时，它通常会将其上传到云端，以便 AI 算法可以重新训练并从中学习。因此，模型在边缘运行的时间越长，模型就会变得越准确，由于可以获得如此多的价值，企业正在迅速采用边缘计算。Gartner 预测，到 2023 年底，50%的大型企业将拥有记录在案的边缘计算战略，而 2021 年这一比例还不到 5%。

图17：边缘 AI 的数据传输

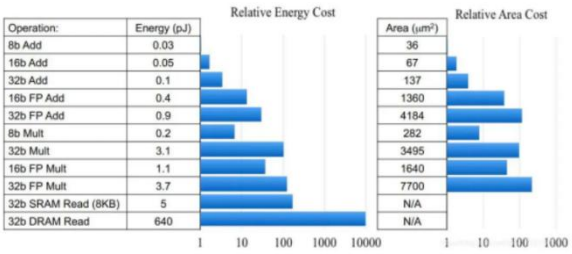

资料来源：NVIDIA，国信证券经济研究所整理

## 缩减优化模型，部署终端设备

通过优化，预估模型算力成本大约会降至原来的 1/4，为模型的边缘部署提供技术基础，目前常见的优化方法有三类：

1）**量化：**量化是模型压缩的一种常用手段，核心思想是将模型参数从高精度转换为低精度，将多 bit 高精度的数（FP32、FP16 等）量化为较少 bit 低精度的数值（INT8、INT4 等），即从浮点到定点数的转换。量化方法可分为**训练时量化**（PTQ，post-training quantization），这种量化方式需要重新训练来缓解量化带来的精度损失；**训练后量化**（QAT，quantization-aware training），在大模型场景上，更青睐于 QAT，因为能够更好的保证性能。量化的优势包括减少内存占用，节省存储空间，降低功耗和占用面积，提升计算速度。

图18：量化可以降低功耗和占用面积




| Operation: | Energy (pJ) | Area (μm²) |
|---|---|---|
| 8b Add | 0.03 | 36 |
| 16b Add | 0.05 | 67 |
| 32b Add | 0.1 | 137 |
| 16b FP Add | 0.4 | 1360 |
| 32b FP Add | 0.9 | 4184 |
| 8b Mult | 0.2 | 282 |
| 32b Mult | 3.1 | 3495 |
| 16b FP Mult | 1.1 | 1640 |
| 32b FP Mult | 3.7 | 7700 |
| 32b SRAM Read (8KB) | 5 | N/A |
| 32b DRAM Read | 640 | N/A |

资料来源：NVIDIA，国信证券经济研究所整理

图19： NVIDIA Turing GPU 体系结构中各种数据类型相对的张量运算吞吐量和带宽减少倍数

| Input | Accumulator | Math Throughput | Bandwidth Reduction |
|---|---|---|---|
| FP32 | FP32 | 1x | 1x |
| FP16 | FP16 | 8x | 2x |
| INT8 | INT32 | 16x | 4x |

资料来源：NVIDIA，国信证券经济研究所整理

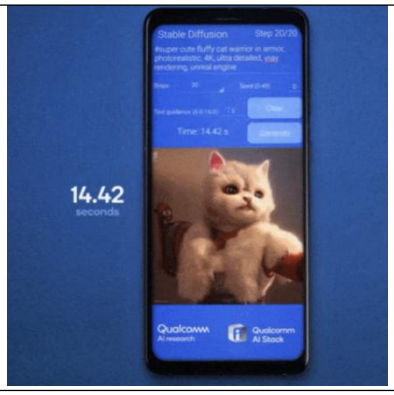
**高通团队采用量化技术等，首次在安卓手机上部署 Stable Diffusion，实现本地运营 15 秒出图，证明了百亿参数级大模型优化后可在终端本地运行的可能。**Stable Diffusion 是一个从文本到图像的生成式 AI 模型，参数达到 11 亿，计算量是智能手机上运行的典型工作负载大小的 10 倍以上，主要限于在云端运行。高通技术团队使用高通 AI 软件栈（Qualcomm AI Stack）执行全栈 AI 优化，使用高通 AI 模型增效工具包(AIMET)对模型进行量化，Hugging Face 的 FP32version1-5 开源模型开始，通过量化、编译和硬件加速进行优化，在搭载 Snapdragon 8 Gen2 移动平台的手机上运行，15 秒内完成了推理，生成一张 512x512 像素的图像。

推理部分是在 Hexagon 处理器上完成的，通过让模型在专用 AI 硬件上高效运行，可消耗更少的内存带宽来节省电量。相比之下，在高通发布 Demo 视频之前，已经有开发者展示了在搭载高通骁龙 865 的 8G RAM 索尼 Xperia 5 II 上运行 Stable Diffusion，生成一张分辨率 512x512 的图像需要 1 个小时。

图20：优化 AI 完全在终端侧高效运行 Stable Diffusion


资料来源：Apple，国信证券经济研究所整理

图21：骁龙 8 Gen2 旗舰芯片组 15 秒出图



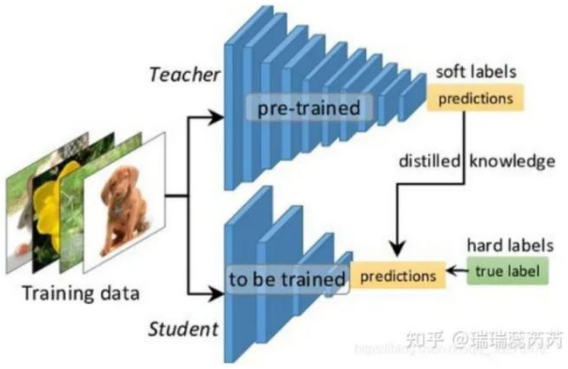

资料来源：Apple，国信证券经济研究所整理

**2）知识蒸馏**（knowledge distillation）：是模型压缩的一种常用的方法，不同于剪枝和量化，知识蒸馏是通过构建一个轻量化的小模型，利用性能更好的大模型的监督信息，来训练这个小模型，以期达到更好的性能和精度。最早是由 Hinton 在 2015 年首次提出并应用在分类任务上面，这个大模型被称之为教师模型，小模型称之为学生模型。来自教师模型输出的监督信息称之为知识，而学生模型学习迁移来自教师模型的监督信息的过程称之为蒸馏。

在子模型场景当中，子模型是完整模型的子集，每个子模型能够独立的训练，学

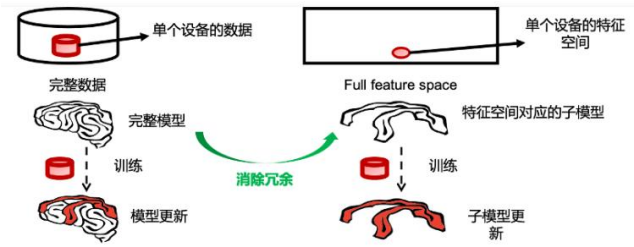
习到某个特定的特征空间的知识。某个设备的数据仅仅占了完整数据集的一部分；映射到特征空间也仅仅是一个区域；使用这个设备的数据集进行训练仅仅更新了完整模型的一部分。因此可以将模型的子集提取出来单独训练，最后整合实现高效的模型更新。

图22：知识蒸馏基本框架




资料来源：NVIDIA，国信证券经济研究所整理

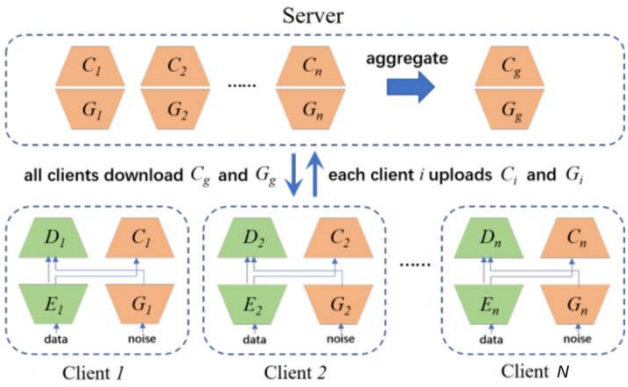
图23：单独训练子模型反哺主模型




资料来源：NVIDIA，国信证券经济研究所整理

**基于知识蒸馏技术，边缘设备除了简单的请求，也可以实现模型更新，反哺集中式数据中心的大模型。**联邦学习最早是谷歌在 2017 年 4 月提出的，可以让数据不离开设备的前提下进行机器学习，且适应性强，保护数据隐私，安全系数高。机器学习模型在现实中的性能表现取决于用来训练它的数据具有多高的相关度，最好的数据就是每天使用的设备。联邦学习会通过服务器发送一部分模型到终端手机，通过几分钟就可以完成训练，然后把训练成果传回服务器。

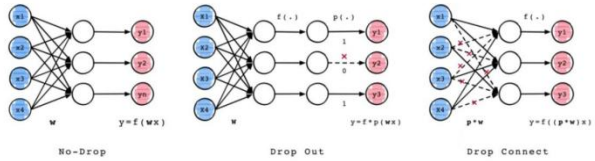
图24：联邦学习的升级版 FedCG




资料来源：量子位，国信证券经济研究所整理

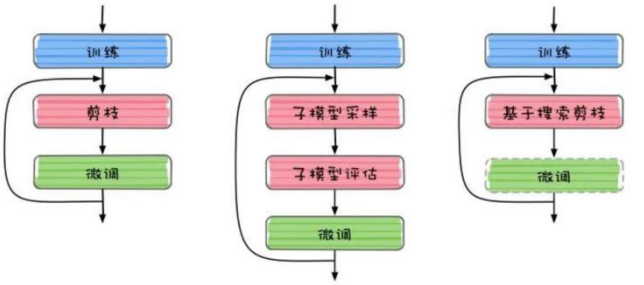
**3）剪枝算法：**过参数化主要是指在训练阶段，在数学上需要进行大量的微分求解，去获取数据中的微小变化，一旦完成迭代式的训练之后，网络模型推理的时候就不需要这么多参数。而剪枝算法正是基于过参数化理论提出的，核心思想是减少网络模型中参数量和计算量，同时尽量保证模型的性能不受影响。主要是分为 Drop Out 和 Drop Connect 两种经典的剪枝算法：Drop Out：随机的将一些神经元的输出置零，称之为神经元剪枝；Drop Connect：随机将部分神经元间的连接 Connect 置零，使得权重连接矩阵变得稀疏。

图25：两种经典剪枝方法




资料来源：CV 技术指南，国信证券经济研究所整理

图26： 剪枝算法流程



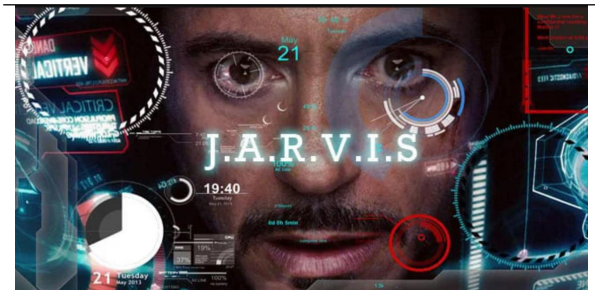

资料来源：CV 技术指南，国信证券经济研究所整理

# “贾维斯”式智能管家，引领全新换机需求

## 大语言模型有望成为复杂 AI 系统的控制中心和交互入口

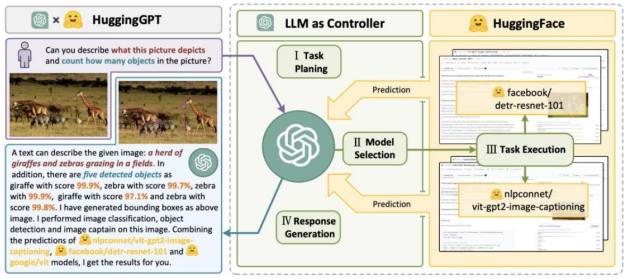
**大模型协作让“贾维斯”式智能管家更进一步。**Jarvis 全称 Just A Rather Very Intelligent System，是漫威宇宙中钢铁侠的 AI 助手，不仅能完成智能家居管理，还能实时监控周围环境、与用户实时沟通、为用户计算最优策略等，这些强大的功能显然这不是一个单独的 AI 模型可以解决的。微软亚洲研究院曾在 Github 上开源过一个叫做 Jarvis 的项目，该系统由 LLM 作为控制器和许多来自 HuggingFace Hub 的 AI 模型作为协作执行者组成，该系统让 LLM 充当控制器来管理现有的 AI 模型，使用语言作为通用接口来调用外部模型，解决实际任务。

图27：钢铁侠和 Jarvis




资料来源：漫威，国信证券经济研究所整理

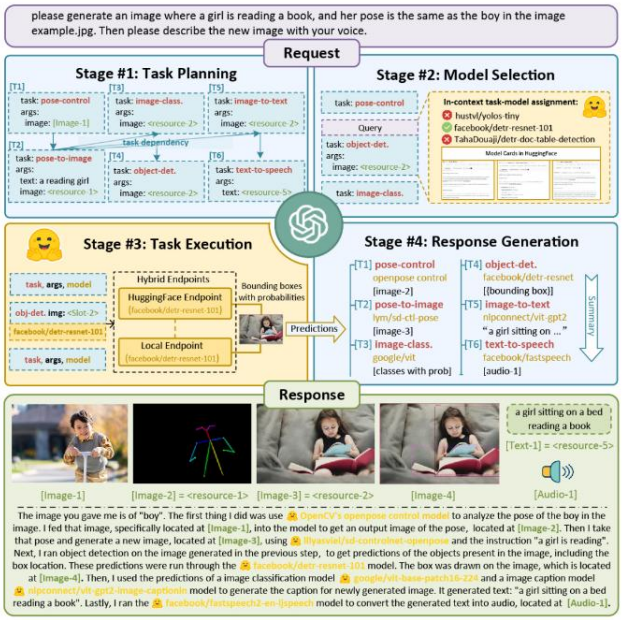
图28：微软亚洲研究院的 Jarvis 项目




资料来源：《HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace》，国信证券经济研究所整理

该系统的具体工作流程包括四个步骤：1）任务规划，使用 ChatGPT 等大语言模型分析用户请求，了解用户意图，并将其拆解成可解决的任务；2）模型选择，为了解决计划的任务，ChatGPT 根据描述选择托管在 Hugging Face 上的 AI 模型；3）任务执行，调用并执行每个选定的模型，并将结果返回给 ChatGPT。4）生成响应，最后使用 ChatGPT 整合所有模型的预测，生成 Response。未来，智能音箱、家用中控屏、甚至于手机、MR 都有可能成为“贾维斯”式管家的交互入口，及时性、可靠性、隐私性或是算力角度，将作为模型协作控制中心的大语言模型部署到边缘设备上必要性越来越强。

图29：Hugging Face AI 模型写作系统四个步骤

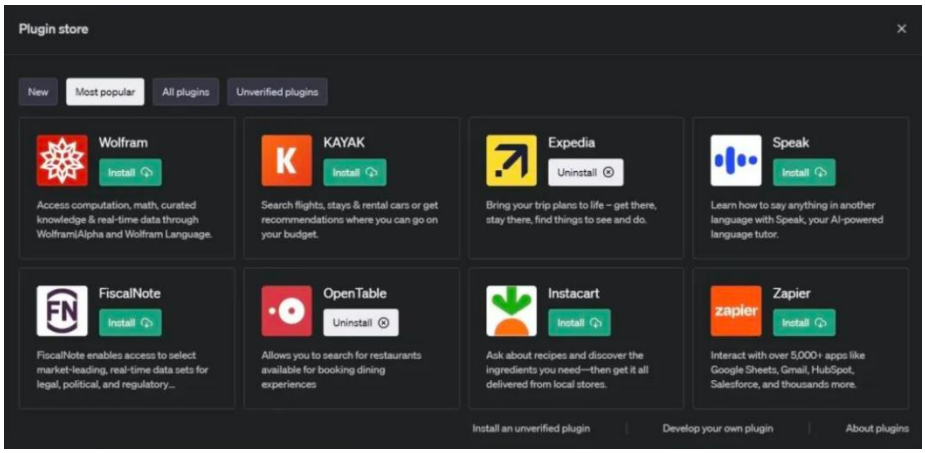



资料来源：《HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace》，国信证券经济研究所整理

**2023 年 5 月第三周，OpenAI 向所有 ChatGPT Plus 用户推出 Beta 版本，有望成为“贾维斯”核心控制中心，打造全新的流量入口和应用生态。**Beta 版本 ChatGPT 支持联网和 70 多种第三方插件，覆盖购物、餐饮、旅行、天气、运算、翻译、分析数据等多种功能。ChatGPT Plugin 的发布为智能助理的出现提供了条件，让语言成为各大模型交互的通用接口。尽管目前尚处于初期，效果不尽如人意，但通过人类语言指挥 ChatGPT 帮自己与各种应用交互仍是令人兴奋的尝试。

随之而出现的，则是用户和 App 提供方的担忧。App 公司普遍担忧 GPT4 太过聪明，接入的 App 不仅害怕数据失去独占优势，还担心 GPT4 通过推理洞悉尚未发现的业务。我们认为，出于数据资源所有权分配以及数据安全的担心将推动终端部署大模型的需求。

图30：Plugin 插件界面

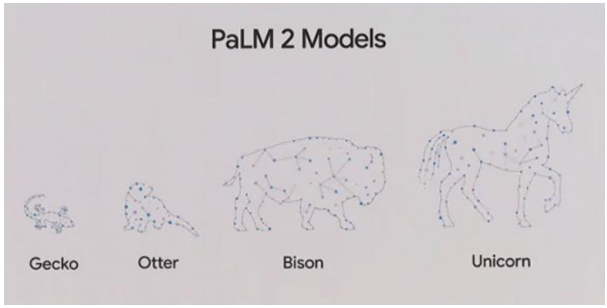



资料来源：36 氪，国信证券经济研究所整理

**各家大厂对终端运行大模型的尝试频频，2023 年 5 月 11 日，Google 在其第 15 届 I/O 开发者大会上发布了 PaLM2，**相比 PaLM 5400 亿参数，它的算法经过优化，使得体积更小，但整体性能更好，计算效率更高，支持 100 多种语言和 20 多种编程语言，支持多模态的 PaLM 2 还能看懂和生成音视频内容。与 ChatGPT 相比，PaLM2 优势在于响应速度更快。谷歌表示，名为 Gemini 的下一代模型将是多模式的，具有突破性的功能，但它仍在接受培训，距离发布还有几个月的时间。

另外，PaLM2 模型从小到大有 4 种版本：“壁虎”（Gecko）、“水獭”（Otter）、“野牛”（Bison）、“独角兽”（Unicorn），实现在不同等级的设备上部署。例如在智能手机上就可以运行规模比较小的 Gecko 模型，让移动端也能拥有大语言模型。Gecko 模型可以在完全离线的情况下在智能手机上运行，它可以在旗舰手机上每秒处理 20 个 token，大约是每秒 16 个单词。谷歌没有明确说明使用了什么硬件来测试，但提到是在“最新的手机上”运行，这证明了与大模型具备类似能力的轻量化版本可以实现本地化部署。

图31：PaLM2 的从小到大的四种版本




资料来源：Google，国信证券经济研究所整理

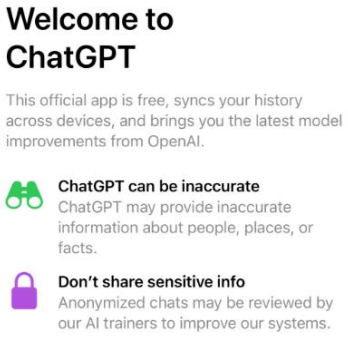
图32：PaLM2 在部分测试中体现出了优异性

| | SOTA | GPT-4 | PaLM | PaLM 2 |
|---|---|---|---|---|
| WinoGrande | 87.5[a] | 87.5[a] (5) | 85.1[b] (5) | **90.9** (5) |
| ARC-C | **96.3**[a] | **96.3**[a] (25) | 88.7[c] (4) | 95.1 (4) |
| DROP | **88.4**[d] | 80.9[a] (3) | 70.8[b] (1) | 85.0 (3) |
| StrategyQA | 81.6[c] | - | 81.6[c] (6) | **90.4** (6) |
| CSQA | **91.2**[e] | - | 80.7[c] (7) | 90.4 (7) |
| XCOPA | 89.9[g] | - | 89.9[g] (4) | **94.4** (4) |
| BB Hard | 65.2[f] | - | 65.2[f] (3) | **78.1** (3) |

资料来源：Google，国信证券经济研究所整理

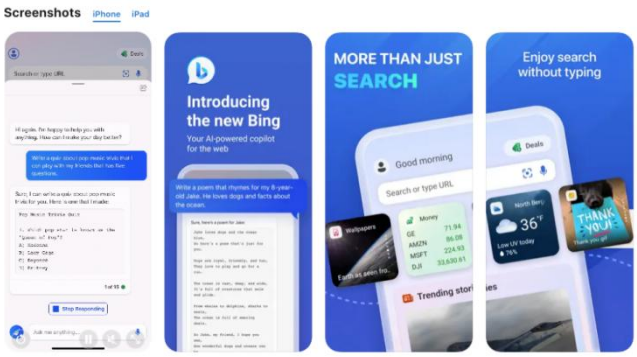
**2023 年 5 月 19 日，OpenAI 在美国区 iOS 商城发布 ChatGPT App，这是用户首次可以在搜索引擎或浏览器之外的移动设备上访问 ChatGPT。**目前 App 可以免费使用，并可同步网页端历史信息。在其欢迎界面上，App 提醒用户 ChatGPT 是有可能提供不准确的信息，并且建议用户不要提供敏感信息，因为匿名信息也可能会被 OpenAI 的训练员用来改善系统。目前 ChatGPT App 仅支持 iPhone8 及更新的机型，支持 iOS 16.1 及更新的系统，切仅支持文字交互模式，不支持多模态的图片或视频输出，也不能调用摄像头，不支持联网、插件功能。ChatGPT 移动端的推出也有望推动谷歌等其他大厂快速跟进，AI 模型全面进入移动端创新阶段。

图33：ChatGPT App 欢迎界面




资料来源：OpenAI，国信证券经济研究所整理

图34：微软 Bing chat 应用

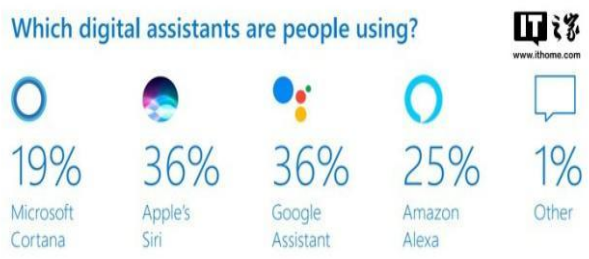



资料来源：Bing，国信证券经济研究所整理

## 当前旗舰机款手机芯片仅可运行优化版十亿参数级大模型

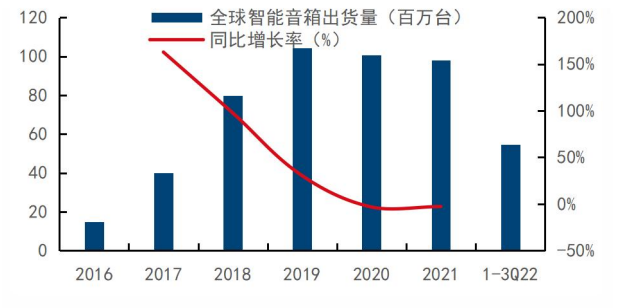
**上一代人工智能程度较低拉低交互体验，阻碍 AIoT 发展。**2011 年 Apple 推出 Siri，使语音助手成为当时人工智能竞赛热门赛道，引发 Google（Google Assistant）、Amazon（Alexa）、微软（Cortana）等科技巨头纷纷加码跟进，抢占 AIoT 控制流量入口。由于通过指挥控制系统进行工作，上一代语音助理仅可以理解有限的问题和请求列表（包含在数据库中的单词列表），如果用户要求虚拟助手做一些代码中没有的事情，机器人会简单地说它无法提供帮助。由于智能化较低，全球语音助理、智能音箱及其他语音交互 AIoT 行业发展经过初期高速成长期后陷入沉寂。

图35：2019 年美国语音助理市场份额




资料来源：IT 之家，微软研究，国信证券经济研究所整理

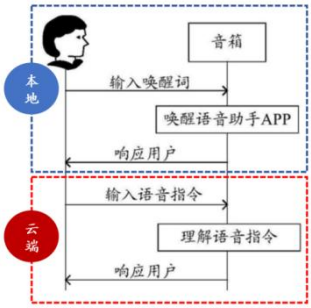
图36：全球智能音箱市场下滑




资料来源：IDC，国信证券经济研究所整理

**以苹果手机的 Siri 为例，目前旗舰机型手机芯片能支持离线唤醒和语音识别。成功唤醒电子设备是实现人机语音交互的基础。**当设备处于待机状态时，需要识别用户输入的语音唤醒信号，如果识别成功则切换到工作状态。目前常见的唤醒方法是通过预设的唤醒参数检测用户的语音输入，唤醒参数如唤醒门限、拾音方向、噪声抑制参数、放大增益等，参数的取值决定了电子设备唤醒率的高低。这通常由一个独立的小芯片，在本地实现。

图37：语音交互过程示意图

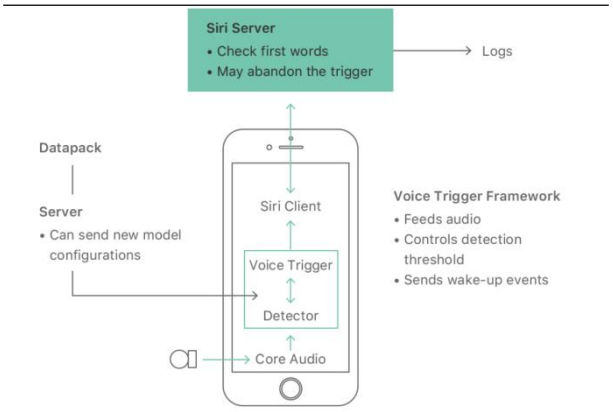



资料来源：集微网，国信证券经济研究所整理

**检测关键词的探测器（Detector）不仅要长时间待机且功耗要足够低到对电池寿命无显著影响，并最小化内存占用和处理器需求。**以 iPhone 的 Siri 为例，iPhone 的 Always on Processor（AOP）是一个小的、低功耗的辅助处理器，即嵌入式运动协处理器。AOP 可以访问麦克风信号，并用自己有限的处理能力运行一个修剪版神经网络模型（DNN）。当分数超过阈值时，运动协处理器唤醒主处理器，主处理器使用较大的 DNN 分析信号。第一个检测器使用 5 层 32 个节点的隐藏单元的 DNN（AOP 运行），第二个检测器使用 5 层 192 个节点的隐藏单元 DNN（主 CPU）。

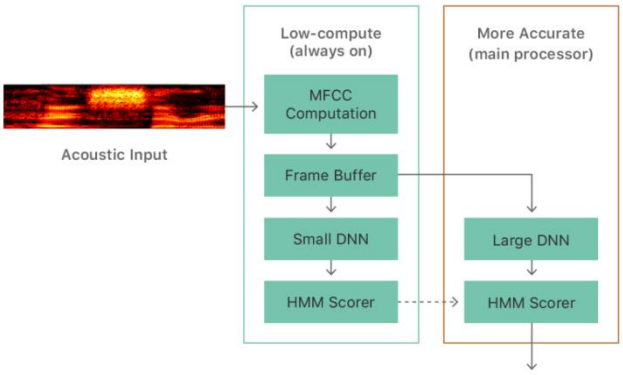
而在 Apple Watch 中，因为电池体积缩小、产品体积缩小，对功耗提出了新的需求，采用的是单通道检测，机器学习模型介于手机上的第一次和第二次检测之间，并仅在手表运动协处理器检测到抬手手势时运行。如果语音内容超过了本地模型的理解范围，数据就会传递到 Siri 服务器，用更复杂的模型识别。

图38：Siri 信号流示意图




资料来源：Apple，国信证券经济研究所整理

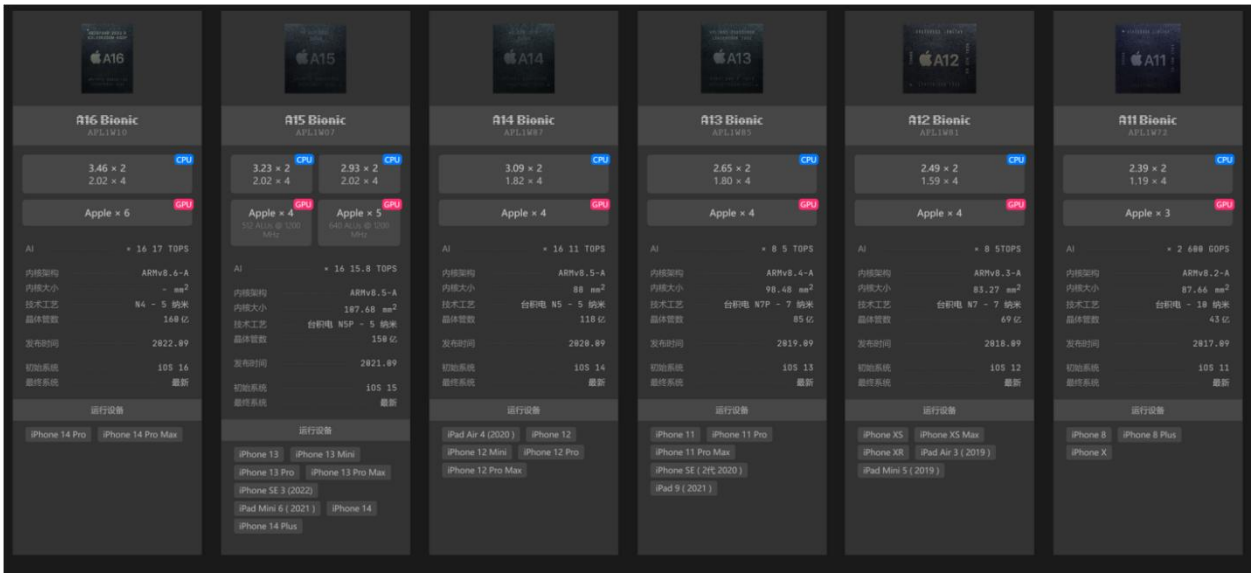
图39：双通检测（AOP 唤醒主 CPU）




资料来源：Apple，国信证券经济研究所整理

**iPhone6 时代，Siri 仅可以离线被“唤醒”，A12 仿生芯片时代，Siri 可以支持部分离线请求。**A11 是苹果首次搭载神经网络引擎处理器单元（Neural Network Processing Unit NPU），但主要是支持面部识别。2018 年苹果推出的 A12 Bionic，采用了台积电 7nm 工艺制程，苹果自研的 Fusion 架构，NPU 从双核直接升级到八核，能够实现每秒 5 万亿次计算。搭载 A12 仿生芯片的 iPhoneXS 首次支持 Siri 离线运行，在不联网的情况下，Siri 可以执行拨打电话、打开特定应用、设置闹钟等请求，也可以实现语音输入等功能，但是无法响应预设内容以外的请求。这说明 10 亿参数以下的 RNN 模型已经完全可以离线运行，但是复杂请求无法实现。

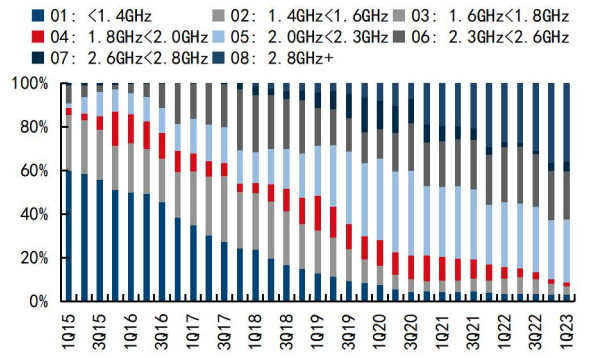
图40：苹果 A11 芯片开始搭载 NPU




资料来源：Apple，国信证券经济研究所整理

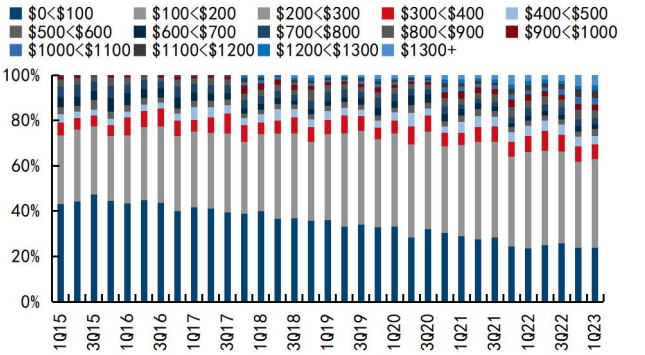
**目前具备运行优化模型能力的终端仅限头部品牌旗舰手机。**根据目前各家公司对于边缘端部署的情况推测，即使经过量化、剪枝、蒸馏等方式优化大模型后，仍然需要旗舰版的手机芯片可以勉强承载运行。假设旗舰机型主处理器频率应在 2.8GHz 以上，或是价格在 1000 美金以上，根据 IDC 数据，1Q23 全球手机销量中主处理器频率超过 2.8GHz 的占比 36%，销售价格在 1000 美金以上的手机销量占比 13%。

图41：全球手机分处理器频率销量占比




资料来源：IDC，国信证券经济研究所整理

图42：全球手机分价格段销量占比




资料来源：IDC，国信证券经济研究所整理

**随着 AIGC 赋能语音助理，AIoT 交互体验升级有望激发终端换机需求。**AI 助手在大语言模型和算力加持下，自然语言理解能力大幅提升，具备了实际生产力后，用户产生使用 AI 助手的需求，从而推动手机换机新周期。此外，智能音箱、全屋智能中控屏、VR/AR/MR 等同样有望成为“贾维斯”的交互入口。

23 年 4 月，脱口秀演员鸟鸟介绍了自己的分身“鸟鸟分鸟”，这个数字分身为阿里训练出来的类 ChatGPT 语音助手，能够模仿她的音色、语气以及文本风格。阿里展示了“鸟鸟分鸟”模型接入智能音箱使用效果，其智能语音交互功能获得颠覆式升级，聊天技能明显升级，AI 有望真正实现对 IoT 赋能。

图43：AIGC 支撑 AI 多模交互

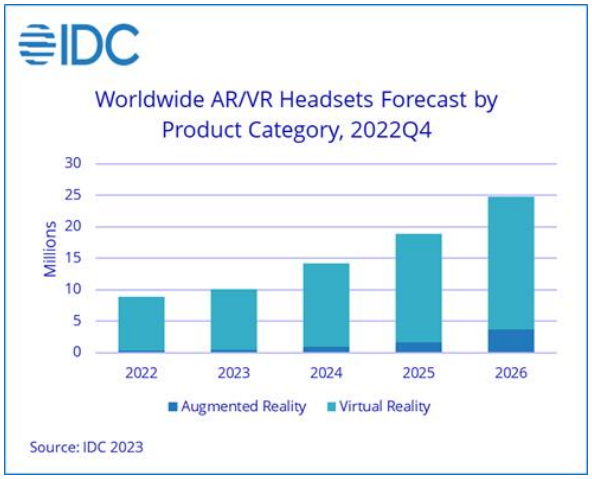
| | 技术 | 阶段 | 作用和目的 |
|---|---|---|---|
| 语音理解 | ASR | 感知阶段 | 将人的语音转换为文本 |
| | NLP | 决策阶段 | 处理并理解文本，以对话能力为核心，为数字人的大脑 |
| | TTS | 表达阶段 | 将需要输出的文本合成为语音 |
| 动作合成 | AI驱动嘴形动作 | 表达阶段 | 建立输入文本到输出音频与输出视觉信息的关联映射，主要是对采集到的文本到语音和嘴形视频（2D）/嘴形动画（3D）的数据进行模型训练，得到相关模型，并智能合成 |
| | AI驱动其他动作 | 表达阶段 | 动作是采用随机策略或者脚本进行预设，需要人工配制描述性的数据或者标签 |

资料来源：腾讯研究院，国信证券经济研究所整理

图44：鸟鸟和类 ChatGPT 模型分身对话


资料来源：阿里巴巴，国信证券经济研究所整理

图45：全球 AR/VR 出货量预测




资料来源：IDC，国信证券经济研究所整理

图46：全球智能家居出货量预测


资料来源：IDC，国信证券经济研究所整理

# 风险提示

**宏观 AI 应用推广不及预期。**AI 技术在应用推广的过程可能面临各种挑战，比如：（1）AI 技术需要更多的时间来研发和调试，而且在应用过程中可能会受到数据质量、资源限制和技术能力等因素的制约；（2）AI 技术的实施需要更多的资源和资金支持；（3）市场竞争可能也会影响企业在 AI 应用推广方面的表现。因此，投资者应审慎评估相关企业的技术实力、资金实力以及管理能力，相关企业的 AI 应用存在推广进度不及预期的风险。

**AI 投资规模低于预期。**尽管 AI 技术在过去几年中受到广泛关注，但 AI 相关领域的企业投资回报并不总是符合预期。部分企业在 AI 领域可能缺乏足够的经验和资源，难以把握市场机会。此外，市场竞争也可能会影响企业的投资力度。因此，存在 AI 领域投资规模低于预期，导致企业相关业务销售收入不及预期的风险。

**AI 服务器渗透率提升低于预期。**虽然 AI 服务器的应用已经较为广泛，但 AI 服务器渗透率提升的速度存在低于预期的风险，这与企业对 AI 技术的投资意愿有关，也可能与市场需求和技术进展的速度有关。

**AI 监管政策收紧。**由于 AI 技术的快速发展和广泛应用，监管机构可能会加强对 AI 技术的监管力度。监管机构可能会制定严格的 AI 技术使用规定，以保障人们的隐私和数据安全，这些监管政策可能会对企业的业务模式和发展战略造成影响。

# 免责声明

## 分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

## 国信证券投资评级

| 类别 | 级别 | 说明 |
|---|---|---|
| 股票投资评级 | 买入 | 股价表现优于市场指数 20%以上 |
| | 增持 | 股价表现优于市场指数 10%-20%之间 |
| | 中性 | 股价表现介于市场指数 ±10%之间 |
| | 卖出 | 股价表现弱于市场指数 10%以上 |
| 行业投资评级 | 超配 | 行业指数表现优于市场指数 10%以上 |
| | 中性 | 行业指数表现介于市场指数 ±10%之间 |
| | 低配 | 行业指数表现弱于市场指数 10%以上 |

## 重要声明



## 证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

# 国信证券经济研究所

**深圳**
深圳市福田区福华一路 125 号国信金融大厦 36 层
邮编：518046　总机：0755-82130833

**上海**
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层
邮编：200135

**北京**
北京西城区金融大街兴盛街 6 号国信证券 9 层
邮编：100032